系牢: 轻量级蒸馏阿拉伯语语音基础模型

Vrunda N. Sukhadia, Shammur Absar Chowdhury

[1]Qatar Computing Research Institute, HBKU, Qatar

sukhadiavrunda@gmail.com, shchowdhury@hbku.edu.qa

Abstract

大型预训练语音模型在下游任务中表现出色，但它们的部署对于资源受限环境来说是不切实际的。本文介绍了 HArnESS，首个以阿拉伯语为中心的自监督语音模型系列，旨在捕捉阿拉伯语发音的特点。通过迭代自我蒸馏，我们训练了大型双语 HArnESS（HL）SSL 模型，并将知识浓缩到压缩的学生模型（HS、HST），保留了特定于阿拉伯语的表现形式。我们使用低秩近似进一步紧凑化教师的离散监督，形成浅层薄模型。我们在阿拉伯语 ASR、说话人情感识别（SER）和方言识别（DID）上评估了 HArnESS，证明其效果优于 HuBERT 和 XLS-R。只需少量微调，HArnESS 就能达到 SOTA 或可比性能，使其成为一种轻量级且强大的实际应用替代方案。我们发布了浓缩模型和研究结果，以支持低资源环境中的负责任的研究与部署。

Index Terms: 自监督模型，蒸馏，基准资源，阿拉伯语下游任务

1. 介绍

自监督学习（SSL）通过从大量无标签语音中提取和学习可迁移的表示，彻底改变了语音处理。这些大型 SSL 模型捕获了丰富的语音表示，使其在广泛的语音相关任务 [1, 2, 3, 4, 5, 6] 中表现出高度有效性。这些模型可以作为特征提取器使用，也可以用少量有标签数据进行微调，在资源有限的情况下显著提升性能。

然而，SSL 模型的泛化能力高度依赖于训练数据的多样性和数量。多语言 SSL 语音模型如 XLS-R[7] 在低资源语言中的表现优于单语种模型，特别是在英语等高资源语言 [8] 训练的模型中表现出色。然而，[9] 的最新研究发现，像 XLS-R 这样的模型倾向于优先处理训练数据丰富的语言，这可能导致代表性不足的语言性能不佳。这表明仅仅依赖多语言模型可能无法确保所有语言的最佳性能，而特定语言的 SSL 模型可能是有效解决这一差距的关键。

阿拉伯语语言在语音处理中提出了独特的挑战，由于其庞大的语言多样性。阿拉伯语遍布 22 个国家，包括超过 20 种彼此难以理解的方言，并受到英语和法语等其他语言的显著影响 [10]。鉴于这种语言复杂性，选择适合阿拉伯语语音任务的 SSL 模型需要能够有效捕捉口语方言细微差别的模型，同时保留文化和音韵多样性。虽然多语言模型很有用，但可能无法完全把握阿拉伯语语音的细节，因此需要专门的阿拉伯语-英语 SSL 模型。然而，训练和部署特定语言的语音 SSL 模型需要大量的计算资源、大规模未标记数据以及较长的训练时间，使其成本高昂且对许多研究者来说难以企及。这些高资源需求阻碍了在资源受限环境中的有效微调和部署。

知识蒸馏已成为在压缩大型模型的同时保持其性能的关键技术。蒸馏使一个更小、更高效的模型（学生）能够从较大的计算成本较高的教师模型中学习，减少了内存使用和推理时间，而不会显著降低性能。早期的工作如 DistillHuBERT [11]，FitHuBERT [12]，DPHuBERT [13]，SKILL [14] 等等 [15, 16] 已将任务无关的知识蒸馏应用于 HuBERT [2]。

在这项研究中，我们介绍了基于 **H**uBERT 的 **阿** abic a**n**d**E**nglish**S**elf-**S**upervised Speech (HArnESS) 模型，这是首个阿拉伯语自监督语音模型系列，共同训练在大规模的阿拉伯语和英语语音

数据上。通过专门针对阿拉伯语和英语进行训练，HArnESS 确保了更好地适应方言变化和语音复杂性，使其成为阿拉伯语语音应用的理想选择。

HArnESS 模型使用迭代自蒸馏进行训练，遵循 HuBERT 架构。这种方法通过在其多个训练迭代过程中使用其前身的预测作为监督信号来提升模型性能。我们利用这种学习范式，并在最初的几次迭代中用 24 层编码器层训练 HArnESS-L，保持其深层架构。在随后的迭代中，我们将知识蒸馏到轻量级的学生模型中，从而产生 HArnESS-S（浅层）和 HArnESS-ST（浅层且窄）架构。此外，我们应用低秩近似来紧凑监督信号，并确保有效的知识转移。

我们评估了 HArnESS-L、HArnESS-S 和 HArnESS-ST 模型在下游任务上的表现，包括自动语音识别（ASR）、说话人情感识别（SER）和方言识别（DID）。我们也比较了这些模型与 HuBERT-large（英语）和 XLS-R（多语言）[7] 模型的性能。

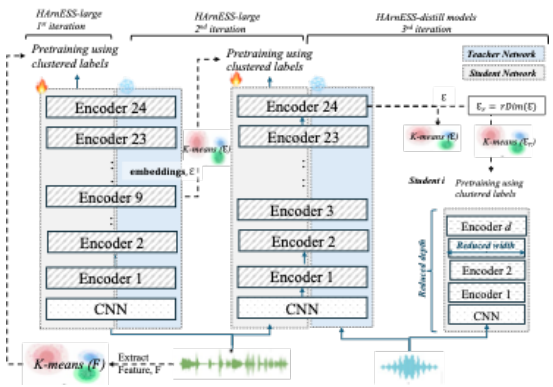


图 1: 概述：通过迭代自蒸馏训练构建 *HArnESS* 模型家族。

因此，我们的关键贡献是：

1. 引入了 HArnESS，这是一个以阿拉伯语为中心的自监督语音模型系列，包括 HArnESS-L（大型）、HArnESS-S（浅层）和 HArnESS-ST（浅层且纤薄）架构。
2. 探索阿拉伯模型压缩和知识蒸馏的迭代自蒸馏范式。
3. 研究紧凑监督信号对模型性能的影响。
4. 公开发布精炼模型（HArnESS-S 和 HArnESS-ST）用于研究。[1]
5. 在内容（ASR）、说话人信息（DID）和副语言（SER）任务上的基准 HArnESS。
6. 发布基准数据集以支持进一步的研究。[1]

2. HArnESS 模型

图 1 提供了 HArnESS 训练管道的概述，该管道遵循类似于 HuBERT 的迭代自蒸馏方法。

训练方案 在每次迭代 i 中，我们使用上一次 $(i-1)$ 迭代模型的伪标签作为监督信号来训练模型 $\boldsymbol{M}_i$。然后该模型学习将随机遮罩的帧分类到这些伪标签中。在前两次迭代中，我们保留相同的架构以增强学生模型的功能抽象。从第三次迭代开始，我们通过蒸馏压缩模型：*(a)* 减少深度（d）以获得更浅的架构；*(b)* 减少宽度（enc_d）以获得更薄的模型和 *(c)* 降低注意力容量（$attn$），通过减少注意力头实现。

模型架构 HArnESS 模型由卷积（CNN）特征提取器和 Transformer 编码器组成。类似于 HuBERT、WavLM [1] 和 Wav2Vec2.0 [17]，CNN 特征提取器包括 7 层时间卷积层。Transformer 编码器层 l 具有嵌入维度 enc_d，并由具有 $attn$ 个头部的多头自注意力（MHA）和位置前馈网络（FFN）组成。

训练目标 我们使用两个交叉熵损失的加权和作为训练目标，这两个损失分别应用于掩码帧和非掩码帧。

权重初始化 我们探索了不同的权重初始化策略以提高收敛性和稳定性。我们试验了随机权重初始化，其中模型权重使用均匀分布的随机数进行初始化，以便在训练开始时引入参数值的多样性。此外，我们还采用了分块平均初始化方法，其中权重通过从前一个模型中平均一组层来初始化。

伪标签生成我们采用 K 均值聚类来为语音帧分配离散标签，利用最后一层的嵌入表示，这些嵌入捕获了模型中最精细的高级表示。[2] 为了提高蒸馏过程中的聚类效率，我们应用主成分分析（PCA）以过滤掉冗余信息，并仅保留最具意义的特征进行表示。对于初始迭代，$i = 0$，我们通过从原始

[1] 为匿名删除了链接。

[2] 我们也探索了从选定层提取的平均嵌入，但未发现显著差异。

语音输入 x 中提取 MFCC 特征来初始化训练，并使用这些特征生成初始伪标签。

3. 实验设置

3.1. 预训练数据

迭代 1 和 2 数据：我们利用了公开的阿拉伯语和英语语音语料库，包括 QASR[18], MGB3[19], LibriSpeech[20], Common Voice（英语和阿拉伯语）[21], GigaSpeech[22] 等。为了使模型具有文化敏感性，我们纳入了来自 15 个阿拉伯语国家的口语内容（从 YouTube 数据中抓取），涵盖了各种方言。最后，我们通过包括速度扰动、SpecAugment、添加混响和噪声等方法来增强所选数据集，以规范并提升模型的鲁棒性。然后我们在这些 **23 千小时**的语音上预训练模型，确保阿拉伯语和英语之间的分布几乎平衡。我们严格排除数据集中的任何官方测试和发展集，以防止从预训练开始的数据泄露并确保可靠的评估。对于训练 k-均值模型，我们使用了 23K 小时数据中的 300 小时子集。

迭代 3 数据：我们的主要目标是开发一个以阿拉伯语为中心的轻量级模型。在这个阶段，我们从 QASR 训练数据集中选择了 **≈1,100 小时** 阿拉伯语数据进行知识蒸馏。为了训练用于生成伪标签的 k-means 模型，我们随机抽取了 Iteration3 数据的 30%（≈300 小时），确保有效聚类的同时保持一定的语言多样性。

3.2. 下游任务和数据

为英语语音 SSL 模型的基准测试已进行了广泛研究，SUPERB[6] 作为评估跨内容识别、说话人信息和副语言任务等任务中 SSL 有效性的关键标准。然而，阿拉伯语语音领域不存在这样的标准化基准。为了填补这一空白，我们引入了阿拉伯语 SSL 模型的基准测试工作，在关键任务上评估性能：ASR 用于内容识别，方言识别（DID）用于说话人信息，以及说话人情感识别（SER）用于副语言分析。对于 ASR，我们在 QASR 数据的一小部分子集（300 小时）上微调 Harness 模型，并在 MGB2 上进行测试。我们通过使用 MGB3 测试集评估该模型来研究这些小型模型在未见过的数据上的泛化能力。对于 SER，我们使用 KSUEmotion[23] 数据集，该数据集从 23 位说话人中收集了 6 种情感类别。数据集分为训练（3.30 小时）、开发（50 分钟）和测试（1 小时）三个部分。[3] 对于 DID，我们使用覆盖 5 个地区方言类别的 ADI5 数据集。详情见表 1。对于下游任务评估，我们选择了广泛使用的指标——ASR 的词错误率（WER），以及 DID 和 SER 的准确度（Acc）。

表 1: 下游任务和数据集统计。*KE*: *KSU* 情感数据集。

数据	**训练（小时）**	**开发 (小时)**	**试验时间（小时）**	
问答系统回顾	300	6.0	–	**自**
MGB2	–	–	9.57	
MGB3	–	–	5.78	
键程	3.30	0.83	1.0	快 惊 生
ADI5	42.90	10.0	10.0	马斯 亚美尼亚

[3] 我们将发布数据分割以确保可重复性。

表 2: 报告的蒸馏模型在 *ASR*、*SER* 和 *DID* 任务上的性能比较。*L*: 大型，*S*: 浅层，*ST*: *S+Thin*。ΔS: 整体结构压缩。*SOTA* *模型是为特定任务设计的大规模模型，并使用大规模/完整对应训练数据进行训练（上限性能）。

模型	自动语音识别（词错误率下降）*(训练：300 小时 QASR 子集)*		SER（准确率 ↑）	DID（准确性提高）
	MGB2	**MGB3**	**KSU 情绪**	**ADI5**
最新技术*	10.24	21.31	83.31%	82.5%
SSL 大型模型				
HuBERT-L (英语)	22.6	51.2	91.92%	64.14%
XLS-R (多语言)	22.60	51.80	73.32%	42.35%
系牢-L (双语：阿拉伯语-英语)	**15.50**	**41.60**	**94.66%**	**84.98%**
压缩使用 ≈ *1000h* 仅阿拉伯语数据				
系牢-S ($\Delta S = 79.4\%$)	*20.20*	*52.80*	*91.15%*	*70.84%*
系泊系统-稳定 ($\Delta S = 93.7\%$)	23.20	58.20	89.02%	69.77%
系牢-ST$^{=}$ ($\Delta S = 93.7\%$)	22.50	55.60	87.34%	61.64%

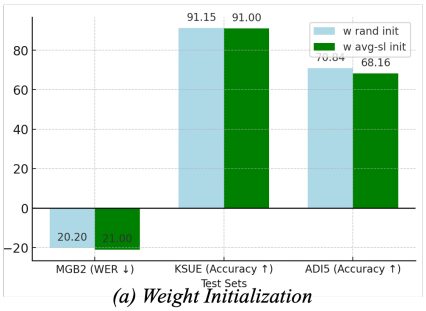
3.3. 预训练训练参数

表 3: *SSL* 模型架构比较。*XR*: *XLS-R*, *HuL*: *HuBERT-Large*, *H-L*: *HArnESS-Large*, *H-S*: *HArnESS-Shallow*, *H-ST*: *HArnESS-Shallow and Thin*。*Dim. dimension*, *Emb.* : *Embedding*。$L*_{emb}$: 来自迭代 i 的模型的第 * 层（例如 *23*）的嵌入。

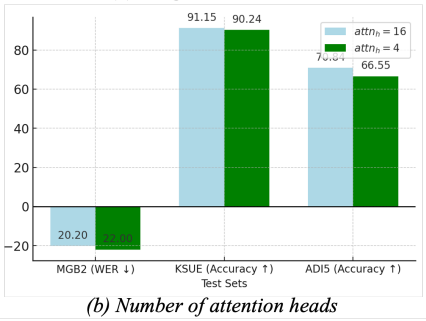
模型	XR	胡磊	H-L	H-S	H-股	H-ST（主成分分析）
监督	–	–	$L9_{\text{emb}}$ $(i=1)$	$L23_{\text{emb}}$ $(i=2)$		PCA($L23_{\text{emb}}$) $(i=2)$
卷积神经网络编码器						
Strides	5, 2, 2, 2, 2, 2, 2					
Kernel Width	10, 3, 3, 3, 3, 2, 2					
Channels	512					
变压器						
Depth (l)	24	24	24	4	4	4
Emb. Dim (d_{emb})	1024	1024	1024	1024	512	512
FFN Dim (d_{ffn})	4096	4096	4096	2048	2048	2048
Attn. Heads (h_{attn})	16	16	16	16	16	16
投影						
Dim. (d_p)	768	768	768	768	768	768
参数						
在 M 中	300	316	316	65	28	28

迭代 1 和 2: 使用 fairseq 代码库对 HArnESS 模型进行了三次迭代训练 [24]。HArnESS-L、HArnESS-S 和 HArnESS-ST 的架构如表 3 所示。在前两次迭代中，我们使用了由 24 层组成的 HArnESS-L 架构。我们在 24 块 H100 GPU 上使用最多每块 GPU 62.5 秒音频数据的批量大小对 HArnESS-L 模型进行了训练，使用的阿拉伯语-英语音频时长为 23k 小时。第一次迭代训练了 50 万步，而第二次迭代训练了 70 万步。在第一次迭代中，我们使用了 39 维的 MFCC 特征，并应用具有 1000 个聚类的 k-means 聚类生成标签。对于第二次迭代，为了获得更好的目标，我们从第一次迭代模型的第 9 层提取潜在表示，并通过使用 1000 个聚类的 k-means 进行聚类来生成新标签。

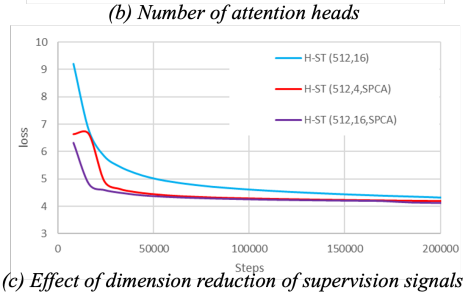
迭代 3: 在第 $i = 3$ 次迭代中，我们使用了 HArnESS-S 和 HArnESS-ST 架构，这些架构分别具有嵌入维度为 1024 和 512 的更浅的四层变压器。我们在 8 块 H100 GPU 上使用每块 GPU 最多 75 秒音频数据的批量大小对 HArnESS-S 和 HArnESS-ST 进行了第三次迭代训练，使用的阿拉伯语音频时长为 1100 小时，并进行了 30 万步的训练。我们从第二次迭代 HArnESS-L 的最后一个变压器层提取特征（聚类数=1000），并使用这些标签来训练这两个模型。因此，可以将这两个模型视为第三次迭代的模型。



(a) Weight Initialization



(b) Number of attention heads



(c) Effect of dimension reduction of supervision signals

图 2: *(a)* 基于学生权重初始化策略的不同下游任务性能；*(b)* 各种 $attn_h$ 的性能；*(c)* 聚类前教师监督信号（*SPCA*）的 *PCA* 效果（emb_d，$attn_h$。

3.4. 下游训练

对于下游任务，我们利用 SSL 模型作为特征提取器，并研究其在捕捉丰富表示方面的有效性。我们将从所有 SSL 层中提取的嵌入进行了平均处理，然后将其作为特征传递给下游网络。

3.4.1. DID 和 SER 架构

对于分类任务，我们选择了一个简单的 CNN，并结合了基于自注意力机制的网络。该网络使用三个连续的时间 CNN 层处理输入的 SSL 特征，核大小为 5，并使用 ReLU 激活函数和 0.4 的 dropout 以实现泛化能力。随后应用自注意力池化，并通过一个 FF 层后传递到输出层。所有层的隐藏维度设置为 80。每个模型使用批大小为 4 进行训练，并运行总共 10K 步。

3.4.2. 自动语音识别

对于 ASR 任务，我们训练了一个编码器-解码器架构，并优化了联合 CTC 和注意力损失。[4] 编码器包含两层 conformer 编码器后跟 2 层 transformer 解码器，每层包括 8 个注意头和 2048 个线性单元。ASR 模型训练了 70 个周期。

4. 结果

与上界比较：SOTA 模型 作为上限，我们展示了阿拉伯语 ASR、DID[25] 和 SER 模型的最新技术水平。表 2 中的结果显示，尽管压缩后的 HArnESS 模型和 HArnESS-L 具有更简单的架构和较短的训练时间，它们在 SER 和 DID 上的表现仍优于最新的模型。这突显了该模型有效捕捉丰富说话人和伴语言信息的能力。对于 ASR 而言，与经过超过 1 万小时方言数据训练的 Fanar ASR[26] 相比，仅使用 300 小时 MSA 微调的 HArnESS 模型分别落后 5（HArnESS-L）和 10（HArnESS-S）个点。

HArnESS-L 对比退出阿拉伯语的 SSLs HArnESS 模型在所有阿拉伯语任务中都优于 HuBERT 和 XLS-R，展示了拥有特定语言模型的重要性。尽管模型规模相似，HArnESS-L 明显受益于特定语言的知识，并且优于一个多语言模型本身。HArnESS-S 和 HArnESS-ST 明显优于 XLS-R，表明压缩后的 HArnESS 也捕捉到了大型模型的抽象表示。

不同结构压缩和设计选择的影响 对于迭代 $i = 3$，我们探索了学生模型中权重初始化对下游任务性能的影响，并观察到没有显著变化（图 2），表明在该训练阶段，初始化作用很小。

通过减少层数，HArnESS-S 较 HArnESS-L 实现了 79.4%的结构压缩，同时在各项任务中保持了强大的性能。它优于多语言和英语模型，证明了其在阿拉伯语语音任务中的效率，尽管进行了压缩。然而，与 HArnESS-L 相比，我们观察到 WER 增加了 4.7，SER 准确率下降了 3.51，DID 准确率下降了 14.4，这表明方言细微差别在较浅

[4] 使用 ESPnet 工具包

的网络中变得不太可区分，使得 DID 成为受影响最大的任务。

进一步将注意力头从 H-S ($attn = 16$) 减少到 H-S*($attn = 4$)，导致了额外的 26.15% 结构压缩（参数数量从 65M 减少到 48M）。虽然 DID 性能受到的影响最大，但 SER 和 WER 的影响仍然很小（图 2）。

通过嵌入维度缩减（表 4），我们观察到相对于 HArnESS-L，在 $\Delta S = 96.52\%$ 压缩比下的性能急剧下降，这突显了过度降维的局限性，显著降低了模型性能。

压缩监督信号如何影响性能？ 为了考察压缩监督信号的影响，我们比较了在第 $i = 3$ 次迭代中知识蒸馏过程中应用 PCA 与未对教师的监督信号进行聚类前应用 PCA 时模型的损失。图 2 显示基于 PCA 的监督收敛速度比其对应方法更快，表明减少监督信号中的冗余可以提高训练效率同时保持有效的知识转移。

表 4: 不同嵌入维度的性能比较。ΔS: 总体结构压缩。

测试集	emb_d**=1024**	emb_d**=512**	emb_d**=256**
MGB2 (WER ↓)	20.2	23.20	22.3
KSUE (Acc ↑)	91.15%	89.02%	79.42%
ADI5 (Acc ↑)	70.84%	69.77%	53.41%
ΔS	70.43%	91.14%	96.52%

5. 结论

在本研究中，我们介绍了 HArnESS，这是首个以阿拉伯语为中心的自监督语音模型家族，旨在捕捉阿拉伯方言的细微差别。通过迭代式自蒸馏方法，我们将大型双语模型的知识转移到浅层（且薄型）的学生模型上，同时保留了特定于阿拉伯语的语音表示。我们在阿拉伯语 ASR、SER 和 DID 任务上的实验表明，HArnESS 达到了最先进的或与现有如 HuBERT 和 XLS-R 等多语言模型相当的结果。轻量级的 HArnESS 也使其成为一个高效但性能有所妥协的选择。我们将提供用于研究目的的轻量级模型及基准数据。

6. References

[1] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1505–1518, 2022.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, pp. 3451–3460, 2021.

[3] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[4] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[5] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 244–250.

[6] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "Superb: Speech processing universal performance benchmark," 2021.

[7] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Proceedings of Interspeech*, 2021.

[8] J. Shi, D. Berrebbi, W. Chen, H.-L. Chung, E.-P. Hu, W. P. Huang, X. Chang, S.-W. Li, A. Mohamed, H. yi Lee, and S. Watanabe, "Ml-superb: Multilingual speech universal performance benchmark," 2023.

[9] E. Storey, N. Harte, and P. Bell, "Language bias in self-supervised learning for automatic speech recognition," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 37–42.

[10] A. Ali, S. Chowdhury, M. Afify, W. El-Hajj, H. Hajj, M. Abbas, A. Hussein, N. Ghneim, M. Abushariah, and A. Alqudah, "Connecting Arabs: bridging the gap in dialectal speech recognition," *Communications of the ACM*, pp. 124–129, 2021.

[11] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7087–7091.

[12] Y. Lee, K. JANG, J. Goo, Y. Jung, and H.-R. Kim, "Fithubert: Going thinner and deeper for knowledge distillation of speech self-supervised learning," in *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*. ISCA, 2022, pp. 3588–3592.

[13] Y. Peng, Y. Sudo, S. Muhammad, and S. Watanabe, "Dphubert: Joint distillation and pruning of self-supervised speech models," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2023, 2023, pp. 62–66.

[14] L. Zampierin, G. B. Hacene, B. Nguyen, and M. Ravanelli, "Skill: Similarity-aware knowledge distillation for speech self-supervised learning," *arXiv preprint arXiv:2402.16830*, 2024.

[15] T. Ashihara, T. Moriya, K. Matsuura, and T. Tanaka, "Deep versus wide: An analysis of student architectures for task-agnostic knowledge distillation of self-supervised speech models," in *23rd Annual Conference of the International Speech Communication Association, INTERSPEECH 2022*, 2022.

[16] R. Wang, Q. Bai, J. Ao, L. Zhou, Z. Xiong, Z. Wei, Y. Zhang, T. Ko, and H. Li, "Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit bert," *Proc. Interspeech 2022*, pp. 1686–1690, 2022.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[18] H. Mubarak, A. Hussein, S. A. Chowdhury, and A. Ali, "QASR: QCRI Aljazeera Speech Resource. A Large Scale Annotated Arabic Speech Corpus," in *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, 2021, pp. 2274–2285.

[19] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic MGB-3," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 316–322.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2020. [Online]. Available: https://arxiv.org/abs/1912.06670

[22] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," 2021. [Online]. Available: https://arxiv.org/abs/2106.06909

[23] A. H. Meftah, M. A. Qamhan, Y. Seddiq, Y. A. Alotaibi, and S. A. Selouani, "King saud university emotions corpus: Construction, analysis, evaluation, and comparison," *IEEE Access*, vol. 9, pp. 54 201–54 219, 2021.

[24] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[25] A. Kulkarni and H. Aldarmaki, "Yet another model for Arabic dialect identification," in *Proceedings of ArabicNLP 2023*, H. Sawaf, S. El-Beltagy, W. Zaghouani, W. Magdy, A. Abdelali, N. Tomeh, I. Abu Farha, N. Habash, S. Khalifa, A. Keleg, H. Haddad, I. Zitouni, K. Mrini, and R. Almatham, Eds. Singapore (Hybrid): Association for Computational Linguistics, Dec. 2023, pp. 435–440. [Online]. Available: https://aclanthology.org/2023.arabicnlp-1.37/

[26] Fanar, U. Abbas, M. S. Ahmad, F. Alam, E. Altinisik, E. Asgari, Y. Boshmaf, S. Boughorbel, S. Chawla, S. Chowdhury, F. Dalvi, K. Darwish, N. Durrani, M. Elfeky, A. Elmagarmid, M. Eltabakh, M. Fatehkia, A. Fragkopoulos, M. Hasanain, M. Hawasly, M. Husaini, S.-G. Jung, J. K. Lucas, W. Magdy, S. Messaoud, A. Mohamed, T. Mohiuddin, B. Mousi, H. Mubarak, A. Musleh, Z. Naeem, M. Ouzzani, D. Popovic, A. Sadeghi, H. T. Sencar, M. Shinoy, O. Sinan, Y. Zhang, A. Ali, Y. E. Kheir, X. Ma, and C. Ruan, "Fanar: An Arabic-Centric Multimodal Generative AI Platform," *arXiv:2501.13944*, 2025. [Online]. Available: https://arxiv.org/abs/2501.13944